УДК 004.9: 575.1

ЗАСТОСУВАННЯ АНАЛІЗУ ГЕННОЇ ОНТОЛОГІЇ ДЛЯ ФОРМУВАННЯ ПІДМНОЖИНИ ЗНАЧУЩИХ ГЕНІВ

І. М. Лях

ДВНЗ «Ужгородський національний університет»,
пл. Народна, 3, Ужгород, 88000, Україна

Охарактеризовано подальший розвиток технології формування підмножин взаємно-експресованих та значущих профілів експресії генів для подальшого їх застосування у системах діагностики на основі даних експресії генів. Запропоновано технологію видалення неінформативних генів за статистичними критеріями із застосуванням аналізу генної онтології, враховуючи кількість генів та характер їх взаємодії. Подано результати практичної реалізації запропонованої технології із застосуванням даних експресії генів пацієнтів, що досліджувалися на різні типи раку. Аналіз отриманих результатів показав високу ефективність запропонованої моделі. Із 19947 профілів експресії генів було виділено 14487 значущих генів, при цьому точність класифікації зразків, що містять як атрибути виділені значущі гени становила 97,6 %. Із 619 зразків, що становили тестову підмножину даних тільки 15 були ідентифіковані некоректно. Представлені дослідження створюють умови для підвищення ефективності гібридної моделі діагностики складних об'єктів на основі даних експресії генів.

Ключові слова: *аналіз генної онтології, експресія генів, тест Фішера, моделювання, класифікація.*

Постановка проблеми. Генна онтологія — це систематизоване визначення генетичних та молекулярних властивостей, яке використовується для опису генетичної інформації та її взаємозв'язків у біологічних системах. Проблема генної онтології полягає в необхідності систематизації та стандартизації генетичної інформації для кращого розуміння функцій генів та їх взаємодій. Одним із ключових завдань є об'єктивне розділення профілів експресії генів на підмножини. У випадку використання нечіткої моделі формування підмножин та гібридної індуктивної моделі кластеризації на основі спектральної кластеризації проблема полягає в розробці ефективних методів формування підмножин інформативних профілів експресії генів. Намагаючись вирішити цю проблему, важливо враховувати оцінку взаємної інформації між відповідними профілями експресії генів, щоб забезпечити точність та релевантність утворених підмножин.

Аналіз останніх досліджень та публікацій. Застосування аналізу генної онтології (GO) для визначення значущих генів має важливе значення в сучасних дослідженнях щодо обробки даних експресії генів. Так, дослідження [1–4] підкреслюють важливість розуміння впливу еволюції GO та її анотацій на інтерпретацію

біологічних експериментів. Виявлено, що послідовність результатів аналізу збагачення, отриманих за допомогою ранніх і більш пізніх версій GO, була низькою. Ця непослідовність пов'язана зі швидким розвитком онтології та анотацій. Було зауважено значне упередження: 58 % анотацій припадають лише на 16 % людських генів. Це дослідження підкреслює необхідність обережності під час інтерпретації аналізів збагачення GO та пропонує переглянути попередні аналізи з використанням найновіших версій GO.

У дослідженні журналу Frontiers in Genetics [5] був розроблений інструмент GeNetOntology, зосереджений на виборі особливостей на основі GO для аналізу генної експресії. Методологія передбачає групування, оцінку та моделювання (G-S-M) для визначення значущих термів GO. Цей підхід інтегрує знання із зовнішніх біологічних ресурсів і використовує алгоритми машинного навчання для завдань класифікації на основі даних експресії генів. GeNetOntology успішно визначив важливі терми онтології, пов'язані з хворобами, що демонструє його потенціал у допомозі генетикам та науковцям у виявленні генів та онтологій, пов'язаних з хворобами, в аналізі транскриптомних даних.

У дослідженні, опублікованому в журналі Genome Biology [6], представлено GeneWalk — метод для ідентифікації окремих генів та їх відповідних функцій для конкретного біологічного контексту. GeneWalk використовує навчання мереж для кількісного вимірювання схожості між генами та їх анотаціями GO. Процес передбачає збірку контекст-специфічної генної мережі, вивчення структури мережі за допомогою випадкових блукань і обчислення значущості схожості між геном і термами GO. GeneWalk виявився ефективним у виявленні відповідних термів GO, перевершуючи альтернативні методи у систематичних порівняннях. Підхід забезпечує швидкий інструмент для генерування гіпотез на основі даних для дослідження функціональних генів.

Вищенаведені дослідження представляють значні кроки в застосуванні аналізу GO для визначення значущих генних підмножин, надаючи цінні відомості про молекулярні механізми, що лежать в основі різних біологічних процесів і хвороб. Еволюція цих методологій відображає динамічну та швидкорозвиваючу сферу геноміки та біоінформатики.

Мета статті — розробка технології формування підмножин значущих профілів експресії генів для подальшого їх застосування у системах діагностики на основі даних експресії генів.

Виклад основного матеріалу дослідження. На сьогодні існує велика кількість баз даних експресії генів різних організмів [7, 8], що постійно поповнюються і містять інформацію щодо цільових генів, які дають змогу ідентифікувати стан відповідного біологічного організму. Проте цільові гени мають складний характер прямих та непрямих взаємодій з іншими вузлами мережі, що визначають їх стан. Одним із методів для формування підмножин значущих генів за статистичними критеріями, який враховує тип об'єкта, що досліджується, що наразі активно використовується у біоінформатиці, є метод на основі аналізу генної онтології (GO). Основною ідеєю методу є використання онтології генів для визначення

генів, які мають вагоме біологічне значення у контексті певного дослідження або експерименту. Ключові аспекти цього методу містять:

– Генна Онтологія (GO): GO є ієрархічною базою даних, яка класифікує генні функції в трьох основних категоріях: біологічні процеси, молекулярні функції та клітинні компоненти. Кожен ген асоціюється з певними термами GO, що описують його роль у клітині.
– Вибір Значущих Генів: в рамках GO-аналізу визначаються гени, які мають високий ступінь асоціації з певними біологічними процесами, молекулярними функціями або клітинними компонентами. Це може бути зроблено за допомогою статистичних тестів для порівняння частоти термів GO серед генів, які є предметом інтересу, з їх частотою у загальній популяції генів.
– Аналіз Збагачення: основна частина аналізу GO полягає у визначенні того, чи є певні GO-терми надмірно представлені (збагачені) серед набору значущих генів. Це може вказувати на те, що ці гени спільно задіяні в певних біологічних процесах або функціях.
– Функціональна Інтерпретація: результати GO-аналізу можуть бути використані для функціональної інтерпретації геномних даних. Наприклад, якщо виявлено, що гени, які асоційовані з певним захворюванням, часто пов'язані з певним біологічним процесом, це може вказувати на ключову роль цього процесу у розвитку захворювання.
– Статистичний Аналіз: для перевірки статистичної значущості збагачення термів GO використовуються різні методи, такі як тест Фішера або аналіз хі-квадрат (Chi-squared).

На рис. 1 зображено блок-схему покрокової процедури застосування аналізу GO для виділення значущих генів на основі анотації GO.

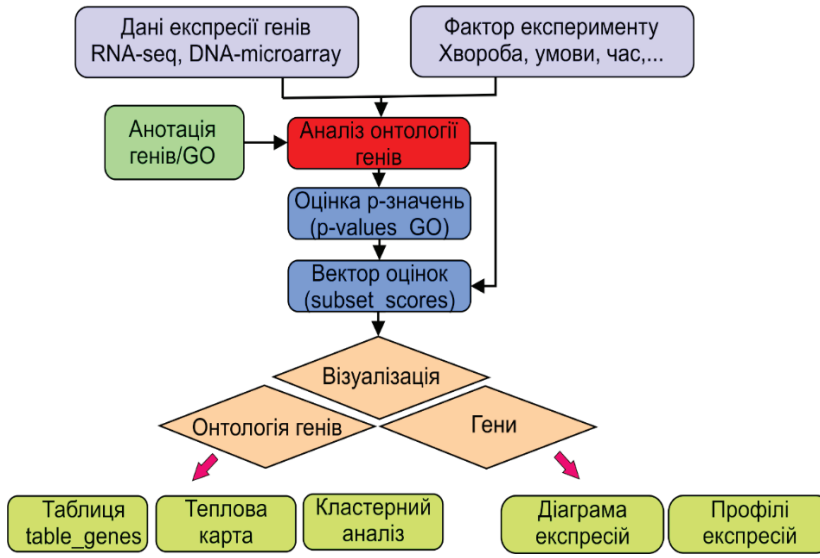

Рис. 1. Блок-схема покрокової процедури застосування аналізу GO для виділення значущих генів на основі анотації GO

Загалом практична реалізація вищенаведеної процедури передбачає наявність таких кроків:

1. Підготовка даних. На цьому етапі формується список генів, що входять до досліджуваних даних. Далі гени анотуються за допомогою існуючої бази даних, яка надає інформацію про їх асоціацію з різними термінами GO.

2. Створення об'єкта генної онтології (GO). На цьому етапі створюється об'єкт онтології, який містить інформацію про всі терми GO та їх взаємозв'язки.

3. Застосування тестової статистики. На цьому етапі статистичний тест застосовується до даних експресії генів, порівнюючи частоту кожного терміна GO у вибраному наборі генів із частотою у фоновому наборі (загальна популяція генів). У межах дослідження на цьому етапі було проведено тести ANOVA та Фішера.

4. Аналіз збагачення термів GO. Ця процедура полягає в оцінці, чи є певні терми GO надмірно представлені (збагачені) серед вибраних генів. На цьому етапі також обчислюється p-значення для кожного терму GO, яке вказує на ймовірність того, що кількість генів з цим термом отримана випадково.

5. Корекція на множинні порівняння. Цей крок зумовлений тим фактом, що, враховуючи велику кількість тестів, які виконуються в аналізі GO, потрібні корекції, щоб уникнути помилкових спрацьовувань. На цьому етапі була проведена корекція p-значень за допомогою тесту Бенджаміні-Хохберга.

6. Інтерпретація та візуалізація результатів. На цьому кроці здійснюється оцінка та аналіз значущих термінів GO, які були ідентифіковані як збагачені серед вибраних генів, аналіз зв'язків між різними термами та створення мережевих діаграм, що відображають біологічні шляхи або процеси. Візуалізація результатів передбачає створення мережевої діаграми найбільш збагачених GO-термів.

7. Формування списку значущих генів, що відповідають найбільш значущим GO-термам. Формування підмножини даних експресії генів, що містять значущі гени як атрибути для подальшого їх аналізу та застосування у системах діагностики стану об'єктів.

Моделювання процесу застосування аналізу GO здійснювалося із використанням даних експресії генів пацієнтів, що досліджувалися на чотири типу ракових захворювань: у 502 пацієнтів була ідентифікована lung squamous cell carcinoma (LUSC), у 541 — lung adenocarcinoma (LUAD), у 542 — kidney renal clear cell carcinoma (KIRC), у 534 — brain lower grade glioma (LGG). Дані, які отримані шляхом застосування методу секвенування молекул RNA (RNA-seq) у межах проєкту The Cancer Genome Atlas (TCGA), є у вільному доступі на інтернет-сторінці проєкту [9]. У початковому стані дані містили 2119 зразків та 19947 генів. Після видалення неекспресованих генів (мають нульову експресію для усіх зразків) кількість генів була зменшена до 19043. Анотація ідентифікаторів генів шляхом порівняння з ідентифікаторами генів, які містяться у базах даних, що відповідають людському організму (був застосований модуль «org.Hs.eg.db»), призвела до зменшення кількості генів до 18930. Були видалені гени, які не анотовані у базі даних.

На наступному кроці до даних застосовувався тест ANOVA для ідентифікації генів з високою диференційною експресією. Аналіз результатів тесту показав, що з 18930 генів 17803 мають високу диференційну експресію (за критерієм p-значення менше ніж 0.01). Наступний етап передбачав створення об'єкта типу topGOdata, який буде містити всі ідентифікатори генів та їх оцінки, анотації GO, ієрархічну структуру GO та іншу інформацію, необхідну для виконання аналізу збагачення відповідних генів.

Перевірка оцінок збагачення передбачає використання тестової статистики: критерій Фішера (на основі обчислення чисел генів) і тест Колмогорова-Смирнова (КС) (на основі обчислення збагачення на основі балів генів). Результати прикладного моделювання тестування збагачення 10 найважливіших онтологій за критерієм Фішера (дані відсортовані за критерієм Фішера) наведено в табл. 1. Аналізуючи дані таблиці, ми робимо такі висновки: велика кількість генів відповідає кожній важливій онтології, і характер зв'язку між онтологією та генами може бути досить складним. На рис. 2 наведено діаграму розподілу десяти найважливіших онтологій. Співвідношення між кількістю генів, включених у відповідну категорію (GO) у вибраному списку генів, до загальної кількості генів у цій категорії відкладено на осі X. Висока частка генів може вказувати на те, що певна категорія GO є важливою в контексті дослідження.

Таблиця 1

Результати застосування тесту на збагачення для десяти найбільш значущих онтологій за тестом Фішера

GO:ID	Терм	Анотовано	Значущі	Фішер	КС
GO:0071840	cellular component organization	6451	6305	< 1e-30	1.1e-14
GO:0070727	cellular macromolecule localization	2443	2408	3.9e-22	1.5e-12
GO:0051649	establishment of localization in cell	2106	2077	1.1e-19	1.3e-09
GO:0051641	cellular localization	3328	3275	9.4e-28	1.2e-13
GO:0051179	localization	5122	4991	4.5e-22	< 1e-30
GO:0044237	cellular metabolic process	9498	9176	1.8e-23	2.6e-07
GO:0044085	cellular component biogenesis	3170	3114	2.5e-23	2.2e-07
GO:0016043	cellular component organization	6241	6096	< 1e-30	3.9e-13
GO:0008152	metabolic process	10813	10417	4.3e-21	4.7e-07
GO:0008104	protein localization	2433	2398	5.9e-22	2.5e-12

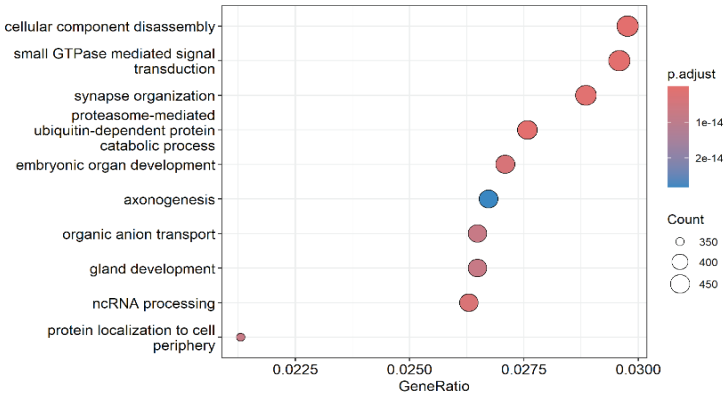

Рис. 2. Діаграма розподілу десяти найбільш значущих онтологій

На рис. 3 зображено спрямований граф взаємодії 20-ти найбільш значущих онтологій (представлені як прямокутники, насиченість кольору визначає ступінь значущості). Всередині прямокутників зазначена також кількість генів, що відповідає онтології. Аналіз діаграми підтверджує складний характер взаємозв'язків між онтологіями та генами. Крім того, результати моделювання показали різницю у результатах при застосуванні тестів Фішера та Колмогорова-Смирнова. З цієї причини під час формування остаточного списку значущих генів використовувалися результати обох тестів.

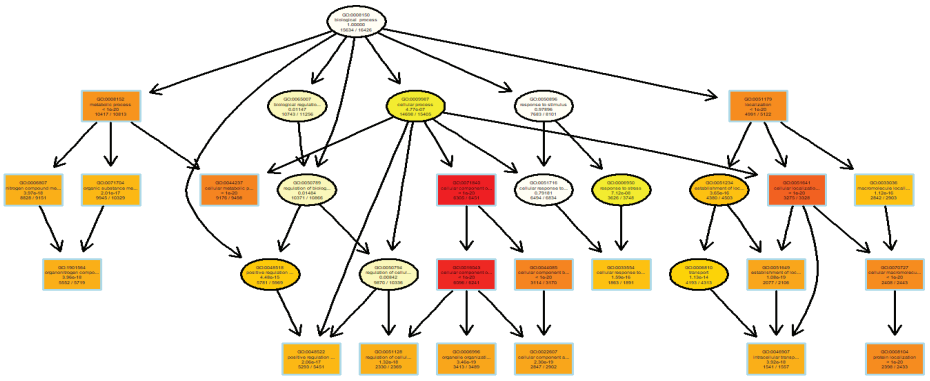

Рис. 3. Мережа взаємозв'язків двадцяти найбільш значущих онтологій

На останньому етапі було створено загальний список важливих генів для обох тестів, виділено унікальні гени та сформовано нові дані, що містять вибрані важливі гени як атрибути. Кількість генів на цій стадії скорочується до 14 488. Важливо зазначити, що з погляду кількості генів отримані результати узгоджуються з результатами моделювання, отриманими при застосуванні нечітких логічних міркувань і систем критеріального аналізу статистики та ентропії.

Оцінка адекватності моделі здійснювалася шляхом застосування класифікатора до сформованих даних. Результати класифікації наведені у табл. 2. Аналіз

отриманих результатів свідчить про високу ефективність методу на основі аналізу GO. Із 619 зразків, що становили тестову підмножину даних, тільки 15 ідентифіковані некоректно. Точність класифікації при цьому становить 97.6 %, що для цього типу даних є досить високою. Високі значення влучності, повторюваності та F1-міри, які визначають якість розподілу зразків в окремі класи, є також досить високими.

Таблиця 2

Результати класифікації даних на основі значущих генів, виділених із застосуванням аналізу GO

Class	Prediction				Precision	Recall	F1	Accuracy
	kirc	lgg	luad	lusc				
kirc	162	1	1	0	0.988	1.000	0.994	97.6 %
lgg	0	159	0	0	1.000	0.994	0.997	
luad	0	0	155	7	0.957	0.957	0.957	
lusc	0	0	6	143	0.960	0.953	0.957	

Але варто зауважити, що кількість генів все ще досить велика. Крім того, формування рішень про статус об'єкта на основі великих баз даних є дуже суб'єктивним. У цьому випадку об'єктивність можна підвищити шляхом розпаралелювання процесу обробки інформації за допомогою аналізу кластеризації або бікластеризації, одночасно використовуючи аналіз GO на кожному рівні для формування списку значущих генів. Це є перспективи подальших досліджень автора.

Висновки. Наведено теоретичні дослідження щодо формування підмножини значущих генів на основі аналізу генної онтології. Наведено покрокову процедуру реалізації методу на основі аналізу GO. Здійснено моделювання процесу виділення значущих генів із застосуванням даних експресії генів пацієнтів, що досліджувалися на чотири типи ракових захворювань. За результатом моделювання із 19947 генів були виділені для подальшого застосування 14487. Результати класифікації підтвердили високу ефективність методу на основі аналізу GO. Із 619 зразків, що становили тестову підмножину даних, тільки 15 ідентифіковані некоректно. Точність класифікації при цьому становить 97.6 %, що для цього типу даних є досить високою. Високі значення влучності, повторюваності та F1-міри, які визначають якість розподілу зразків в окремі класи, є також досить високими.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Tomczak A., Mortensen J. M., Winnenburg R. et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci Rep*. 2018. 8. 5115. Doi: https://doi.org/10.1038/s41598-018-23395-2.
2. Latrille T., Rodrigue N., Lartillot N. Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *PNAS* 2023. Vol. 120. No. 11. March 10, 2023. DOI: https://doi.org/10.1073/pnas.2214977120.

3. Sharn H. O., Singh D. B., Yadav. P. K., Gautam B., Kumar V., Singh S. Genome annotation and comparative functional analysis of genomic islands in Bordetella pertussis Tohama I, Bordetella parapertussis 12822, and Bordetella bronchiseptica RB50 genomes. *Network Modeling Analysis in Health Informatics and Bioinformatics*. 2023. 12 (1), art. no. 23. DOI: 10.1007/s13721-023-00418-1.
4. Huang H., Song J., Feng Y., Zheng L., Chen Y., Luo K. Genome-Wide Identification and Expression Analysis of the SHI-Related Sequence Family in Cassava. *Genes*. 2023. 14 (4):870. Doi: https://doi.org/10.3390/genes14040870.
5. Ersoz N. S., Bakir-Gungor B., Yousef M. GeNetOntology: identifying affected gene ontology terms via grouping, scoring, and modeling of gene expression data utilizing biological knowledge-based machine learning. *Frontiers in Genetics*. 2023. 14, art. no. 1139082.
6. Ietswaart R., Gyori B. M., Bachman J. A. et al. GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biol*. 2021. 22. 55. Doi: https://doi.org/10.1186/s13059-021-02264-8.
7. ArrayExpress - Functional Genomics Data. URL: https://www.ebi.ac.uk/biostudies/arrayexpress.
8. Gene Expression Omnibus – GEO. URL: https://www.ncbi.nlm.nih.gov/geo/.
9. The Cancer Genome Atlas Program – TCGA. URL: https://www.cancer.gov/ccg/research/genome-sequencing/tcga.

REFERENCES

1. Tomczak, A., Mortensen, J. M., & Winnenburg, R. et al. (2018). Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations: Sci Rep, 8, 5115. DOI: https://doi.org/10.1038/s41598-018-23395-2 (in English).
2. Latrille, T., Rodrigue, N., & Lartillot, N. (March 10, 2023). Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale: PNAS 2023, 120, 11. DOI: https://doi.org/10.1073/pnas.2214977120 (in English).
3. Sharn, H. O., Singh, D. B., Yadav, P. K., Gautam, B., Kumar, V., & Singh, S. (2023). Genome annotation and comparative functional analysis of genomic islands in Bordetella pertussis Tohama I, Bordetella parapertussis 12822, and Bordetella bronchiseptica RB50 genomes: *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12 (1), 23. DOI: 10.1007/s13721-023-00418-1 (in English).
4. Huang, H., Song, J., Feng, Y., Zheng, L., Chen, Y., & Luo, K. (2023). Genome-Wide Identification and Expression Analysis of the SHI-Related Sequence Family in Cassava: Genes, 14 (4), 870. DOI: https://doi.org/10.3390/genes14040870 (in English).
5. Ersoz, N. S., Bakir-Gungor, B., & Yousef, M. (2023). GeNetOntology: identifying affected gene ontology terms via grouping, scoring, and modeling of gene expression data utilizing biological knowledge-based machine learning: *Frontiers in Genetics*, 14, 1139082 (in English).
6. Ietswaart, R., Gyori, B. M., & Bachman, J. A. et al. (2021). GeneWalk identifies relevant gene functions for a biological context using network representation learning: Genome Biol, 22, 55. DOI: https://doi.org/10.1186/s13059-021-02264-8 (in English).
7. ArrayExpress - Functional Genomics Data. Retrieved from https://www.ebi.ac.uk/biostudies/arrayexpress (in English).

8. Gene Expression Omnibus – GEO. Retrieved from https://www.ncbi.nlm.nih.gov/geo/ (in English).
9. The Cancer Genome Atlas Program – TCGA. Retrieved from https://www.cancer.gov/ccg/research/genome-sequencing/tcga (in English).

doi: 10.32403/1998-6912-2023-2-67-136-144

APPLICATION OF GENE ONTOLOGY ANALYSIS FOR THE FORMATION OF A SUBSET OF SIGNIFICANT GENES

I. M. Liakh

Uzhhorod National University,
3, Narodna Square, Uzhhorod, 88000, Ukraine
igor.lyah@uzhnu.edu.ua

The development of the technology of forming subsets of mutually expressed and significant gene expression profiles for their further use in diagnostic systems based on gene expression data is characterized. A technology for removing uninformative genes based on statistical criteria using gene ontology analysis, taking into account the number of genes and the nature of their interaction, is proposed. The results of the practical implementation of the proposed technology are presented using gene expression data from patients tested for various types of cancer. The analysis of the obtained results shows the high efficiency of the proposed model. Out of 19947 gene expression profiles, 14487 significant genes are identified, and the classification accuracy of samples containing the identified significant genes as attributes was 97.6 %. Of the 619 samples that made up the test data subset, only 15 are identified incorrectly. The presented research creates conditions for improving the efficiency of a hybrid model for diagnosing complex objects based on gene expression data. The key aspects of gene ontology are considered, namely: gene ontology (GO); selection of significant genes; enrichment analysis; functional interpretation; statistical analysis. A flowchart of a step-by-step procedure for applying GO analysis to select significant genes is also presented.

The enrichment estimates are verified using test statistics: Fisher's criterion and Kolmogorov-Smirnov test and a common list of important genes for both tests is created, unique genes are identified and new data containing selected important genes as attributes are generated. The obtained results are consistent with the modelling results obtained by applying fuzzy logic reasoning and criterion analysis systems for statistics and entropy, and the adequacy of the model is assessed by applying a classifier to the generated data.

Keywords: *gene ontology analysis, gene expression, Fisher's test, modelling, classification.*

Стаття надійшла до редакції 25.08.2023.
Received 25.08.2023.